Design and Evaluation of Gesture Interface of an Immersive Walk-through Application for Exploring Cyberspace

Rieko Kadobayashi Kazushi Nishimoto Kenji Mase
ATR Media Integration & Communications Research Laboratories
2-2, Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-02, Japan
{rieko/knishi/mase}@mic.atr.co.jp

Abstract

We have developed an application of a full-body, non-contact gesture interface for exploring cyberspace that provides immersive walk-through and information accessing capabilities. The VisTA-walk system is designed for use at the museum exhibitions where easy, durable, and pleasurable user interface is preferable. It uses a large projection screen for immersive cyberspace presentation. VisTA-walk allows the user to walk through "virtual villages" by taking physical steps and to retrieve information on objects displayed in the scene by pointing at them.

We carefully designed the interface of VisTA-walk with a simple video-based gesture recognition module, providing it with a minimal but comprehensive set of gestures as its vocabulary. The mouse and gesture-based interfaces are compared through subjective experimentation on walk-through capability for ease of use and degree of immersiveness. The immersiveness achieved by the combined use of the large screen and gesture interaction is comprehensively evaluated.

1. Introduction

Virtual museums are one of the most effective applications of cyberspace, providing virtual, visual and information spaces. Video wall displays are often equipped for immersive presentation of exhibits and are also useful in presenting to groups of people. However, the user interface currently available for cyberspace, e.g. keyboard, mouse, data glove and HMD, are not only cumbersome for the general public but also have not been specifically designed for such exhibits. A non-contact interface is especially desirable for immersive exhibits that use large video walls for displaying cyberspace.

We have developed an application system for a virtual museum exhibit using the pfinder [12], which is a full-body and non-contact gesture recognition module, as an interface for exploring cyberspace. This interface provides ease to use, no anxiety about breaking the system, system durability, and pleasurable physical actions. The system is named VisTA-walk, and is an interactive simulation tool for handling archaeological data. It has a gesture interface for museum visitors and provides immersive virtual walk-through and information-accessing capabilities. VisTA-walk uses a large projection screen (170 inches) for immersive cyberspace presentations. With gesture interactions, it allows users to walk through virtually reconstructed villages with physical steps and to get information on objects in the scene by using pointing gestures.

Full-body gesture interfaces have been explored in various applications, such as environments of artificial life[8], music and dance instruments[1, 2], video games[3, 7], play spaces[16] and computer assisted performance[11]. The VisTA-walk system uses the pfinder as its gesture interface and is similar to the DOOM game application[13] in that both systems provide a virtual walk-through and interaction with the object in the virtual world. We carefully designed the interface of VisTA-walk, providing it with a minimal but comprehensive set of gestures as its vocabulary. We separately assigned stepping and hand stretching gestures to the walk-through and the object pointing, respectively. However, we left voice commands, which are used in DOOM for indiscriminate object pointing, for work on future multimodal interactions with a richer set of vocabulary.

In this paper, mouse and gesture-based interfaces for walk-though are compared through subjective experimentation for ease of use and degree of immersiveness. The rest of this paper is organized as follows. Section 2 provides the motivation of this research and an overview of the VisTA and VisTA-walk systems. Section 3 discusses the usability experiments of their user interfaces for museum exhibits. Section 4 concludes this paper.

2 VisTA and VisTA-walk as Museum Exhibits

2.1 Testbeds for Future Museums

Generally, museums are places where artifacts are displayed and people study and enjoy history and art. However, it is difficult for traditional exhibits to convey a rich knowledge of experts to all types of museum visitors who may have different ages, interests, and knowledge levels. Novel ways of presentation, such as using virtual reality technology, is necessary for future museums[6].

Moreover, traditional exhibits cannot respond to visitor requests for detailed information on artifacts and for access to basic data. Access to basic data should be available through intuitive and easy operations.

Nevertheless, it is important to consider museums as institutions that bind experts and members of the general public, who do not have expert knowledge. In order to attract the general public, the exhibition should be fun from the start and lead to highly interactive sessions that allow the visitor to explore the vast knowledge database of experts.

Thus, we have proposed Meta-Museum as a new concept of the museum institution[5, 9] and have developed the VisTA and VisTA-walk systems based on the concept. They simulate the transition process of an ancient village[14]. Users can visualize the transition process through real-time 3D computer graphics after they interactively set the value of each building's lifetime. Users intuitively know the ancient landscape of the site because they can walk through the reconstructed 3D CG village. The systems provide intuitive information access through the selection of objects such as buildings in the 3D CG scene. Moreover, the gesture interaction of the VisTA-walk system give users more pleasure by letting them use natural physical actions and relieve their anxieties about using a delicate computer system.

2.2 The VisTA system

VisTA consists of a module for managing a historical site database, a space-time simulation module, a Web browser and documents in HTML format, and viewers. With the database management module, a user can store 3D land models, locations of excavated buildings, and 3D building models as basic data and input and/or modify hypotheses on space-time transitions from an editing window. The space-time simulation module allows users to walk through the visualized ancient village in 3D CG by using two viewers that provide scenes from different viewpoints as well as to freely set the simulated year. The simulation and visualization module is implemented in C++ using the Open Inventor library. The Web browser is connected to the main object in the viewers, and if this object is selected it shows the relevant HTML documents[1]. Fig. 1 shows the console window, and a part of Fig. 2 shows the internal structure of VisTA.

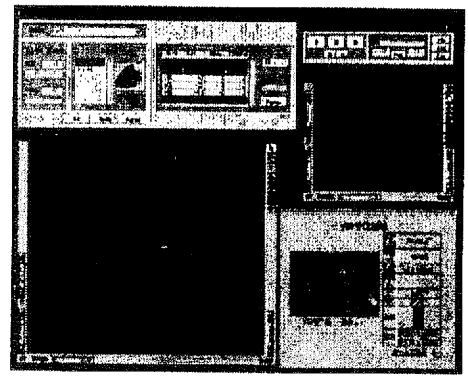

Figure 1. User interface of VisTA

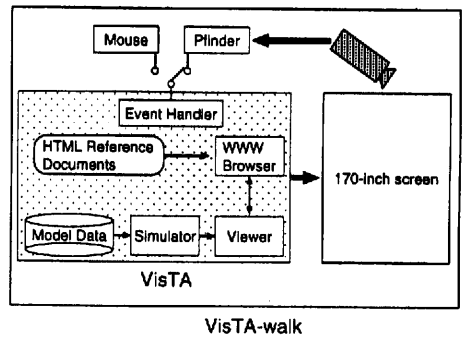

Figure 2. Block diagram of the systems

We displayed the VisTA system at the Ori-Amu Museum in Osaka from November 4 to December 3, 1996 to evaluate the effectiveness of the system for non-experts. For the user interface, this version provided only a mouse as the input device to let users concentrate on experiencing the contents. After analyzing questionnaires completed by visitors, we concluded that even the simplified mouse user interface is difficult for those unfamiliar with PCs and should be improved, even though the system greatly helped visitors understand the contents of the exhibition.

2.3 Overview of VisTA-walk

VisTA-walk has a gesture interface for input and a 170-inch screen as its output interface in addition to the original mouse and keyboard interfaces of VisTA (Fig. 2).

The camera installed on top of the screen captures the image of the user standing in front of the screen. This image

[1] It is easy to convert this part of the system to VRML.

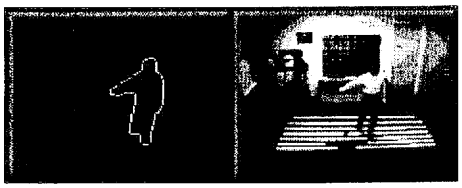

Figure 3. Example image captured by pfinder

is sent to the pfinder program[12], which detects the gestures and position of the user. Pfinder first extracts the area of the user from the captured image and recognizes the head, hands, body, and feet of the user. It then recognizes gestures like standing, sitting, and raising hand(s) and the position of the user. Figure 3 shows an example of the captured image and the edge of the user extracted by pfinder.

Gestures recognized through the process are interpreted as commands to either walk through the virtual village displayed on the screen, change viewpoints, or select a building and get information[10]. The following are major interactions of VisTA: (i) walking through a virtual space for exploration, and (ii) pointing at an object in a virtual space to get reference records. Users are allowed to use different parts of the body for these two independent interactions. Based on this principle, a plausible approach to interaction design is for the standing position to control walking-through and for the hand gestures to control pointing. Table 1 lists the gesture commands.

2.3.1 Gestures for Walk-through

How can we assign a standing position to the control events of the walk-through system? There are at least three choices of control schemes: mouse, joystick[13], or steering wheel and accelerator. The positioning action of a mouse is that of a locator-type device, while a joystick and a steering wheel are valuator-type devices. According to the terminology proposed in [15], a valuator-type device is classified to the "Isotonic Rate" control scheme. We employ a steering wheel-like control scheme for gesture interaction in the VisTA-walk system.

Similar to the movement of a steering wheel, stepping aside from the neutral position is used to steer the direction of walking. Stepping forward or backward is assigned to an accelerator with which a user moves the viewing position forward or backward in the virtual environment. In order to stop at a desired point in a scene, a user must return to the neutral position.

The speed of steering and walking is proportional to the stepping distance from the neutral position, and the ratio of speed to distance affects the assessment of usability. If the speed is too fast, a user can quickly arrive within the proximity of a destination point, but it is difficult to stop at the exact desired point. On the other hand, if the speed is too slow, a user may have to step far from the neutral point to quickly arrive at a point; stopping at the desired point necessitates moving back by several steps, and this causes inaccuracy.

Table 1. List of gesture commands

Gesture	Command
stand on the neutral position	stop
step forward	move forward
step backward	move backward
step right	rotate right
step left	rotate left
raise a hand	select a building
raise both hands	look up
crouch	lower viewpoint

The reason we chose not to use the mouse control scheme is that the area in which the user can move is limited; furthermore, the user cannot warp over the position without creating sensor events, which may be often desired in the mouse control scheme. In joystick control scheme, the position is used for the velocity vector. This would be a good scheme if we had another way to change the orientation, for example by using a horizontal body rotation recognition algorithm. In any case, it would require a much larger vocabulary of recognized gestures or require more dimensions in the recognized space. The steering wheel control scheme makes it possible to come face to face with the target building by only using walking gestures.

The gestures of raising both hands and crouching are interpreted as commands to change viewpoint. A user raises both hands to look up and crouch to lower the viewing position. These changes in the viewing elevation and the viewing position facilitate immersive sensations for a user.

2.3.2 Gestures for Pointing at Objects

The current interpretation of pointing gestures are a "left one" or a "right one". The VisTA-Walk gesture interpreter uses the output recognition results of pfinder, e.g., "stretch left/right hand" or "raise left/right hand". The interpreter chooses an object on the left (right) side when it detects a "left (right) one" request from pfinder outputs. Fig. 4 depicts how a user uses VisTA-walk. In this case, the user raises her left hand to select the house and get information about it in the right top window. That is, raising one hand is interpreted as an object selection command.

The pointing actions recognized by pfinder are insufficient to locate a particular object among many, and this is due to the camera position. Pointing and stretching a

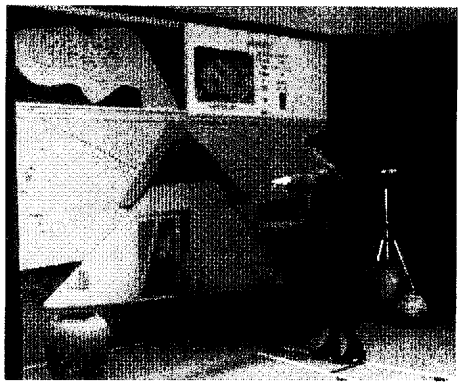

Figure 4. A snapshot of a user in VisTA-walk

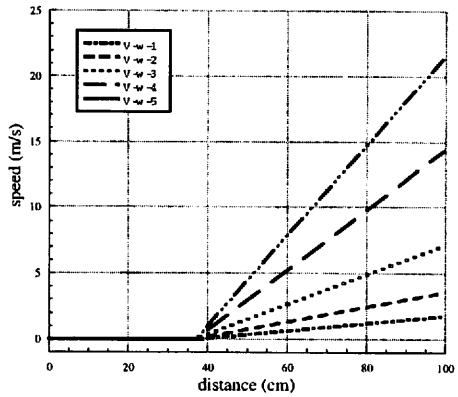

Figure 5. Mapping between distance and speed

hand toward the screen is hard to detect because the camera is on top of the screen. Stereoscopic multiple camera arrangement is necessary to get complete 3-D information about hand and body gestures[4], and such an arrangement is presently impractical because many cameras would be necessary to cover the complete area of a user in motion.

3 Experiment of Usability

In order to evaluate whether the gesture interface is effective as a user interface for systems used in museum exhibits, we executed subjective experiments on VisTA, VisTA-walk and VisTA170, which was prepared for the experiment. VisTA170 uses a mouse as its input device, which is set on the stand located in front of the screen, and a 170-inch screen as its output device.

The subjects were five researchers, six secretaries, five undergraduate students, and four expert users of VisTA and VisTA-walk. Researchers are fully practiced in operating PCs and mouse, and secretaries are rather practiced in this operation while students are not so good at handling a mouse. Except for the expert users, none of the subjects has ever used VisTA, VisTA-walk, or VisTA170 before.

The learning effect of these three kinds of systems was compensated by execution in different orders for different subjects. As for VisTA-walk, in order to observe the change in operability according to the change in mapping between the distance of the user's movement and speed in the virtual space, we tested five different speeds. The slowest VisTA-walk is called VisTA-walk-1 (V-w-1), and the fastest VisTA-walk is called VisTA-walk-5 (V-w-5). The linear mappings between distance of movement and speed are shown in Fig. 5. The mappings of rotational movement are done in similar way.

The task for subjects is to walk through the course (Fig. 6) from start to goal as fast as possible while passing three

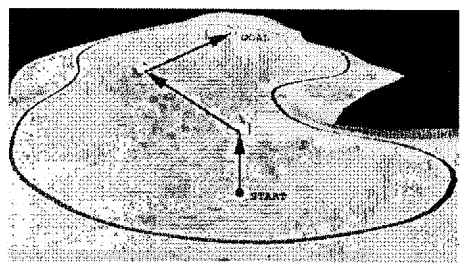

Figure 6. Experimental course

houses. To check if they pass the specified houses in the correct order, we ordered subjects to stop at the center of each house.

For each subject, this task is repeated three times for VisTA, each of the five different VisTA-walk speeds, and VisTA170, i.e. seven systems in total, and we measured time and distance. In addition, we asked subjects to fill out a questionnaire consisting of a five-grade evaluation and a place for free comments after they finished one experiment. The results of subjective evaluations and measured values of time and distance are shown in Figs. 7, 8, and 9, respectively.

3.1 Evaluation 1: Realistic sensation and operational ease

From Fig. 7, we can see that VisTA has a less realistic sensation than VisTA170 and the five different VisTA-walk speeds. However, compared with VisTA170, VisTA is easier to operate. This is supported by the fact that average time and average distance of VisTA170 are longer than those of VisTA (Fig. 8 and 9). In particular, the difference of distance is significant by the t-test. Accordingly, we can say

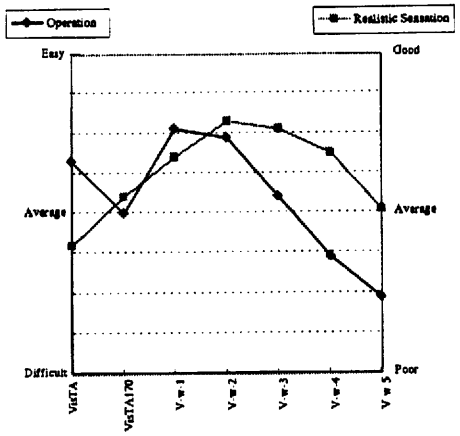

Figure 7. Subjective rating of the seven systems (average of all subjects)

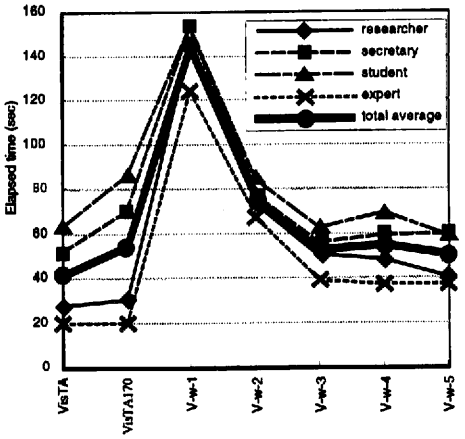

Figure 8. Average elapsed time

that a large screen can give a more realistic sensation while it decreases operational ease when it is used with a traditional mouse interface. There is no significant difference between VisTA and V-w-3 or V-w-4 in terms of average elapsed time. This suggests that the gesture interface is equal to a mouse in operational efficiency.

Consequently, it is not sufficient to simply enlarge the screen but necessary to apply a suitable interface to provide both a realistic sensation and an easy-to-use interface. It is also clear that a gesture interface can provide an operational efficiency equal to that of VisTA while providing a more realistic sensation than VisTA. We can thus conclude that a gesture interface is suitable for systems that use a large screen for immersive presentations in museum exhibits.

3.2 Evaluation 2: Speed and operational ease

What is the appropriate mapping between the amount of movement in real space and the speed in virtual space when a gesture interface is used for walking through the virtual space? The time needed to travel the course can be theoretically reduced to 50%, 25%, 12.5%, and 8.3% in the V-w-n (n = 2,3,4,5) systems compared with V-w-1. However, the experimental results are different. Though travel time is reduced to about 52% in V-w-2, it is reduced at most to about 36%, 38%, and 35% in V-w-3, V-w-4, and V-w-5, respectively. As for the migration distance, it is almost equal to the theoretical distance of the course in V-w-1 and V-w-2. It is longer than the theoretical distance in V-w-3, V-w-4, and V-w-5 (Fig. 9).

To explain this situation, we show an example of typical movements of a subject in Fig. 10. The horizontal axis indicates the elapsed time from start to goal, and the vertical axis indicates the distance in a straight line to the target house. Since there are three target houses in the experimental course, there are three points where the distance is almost equal to zero meter[2]. The distance to each target decreases linearly in V-w-2. In contrast, increase and decrease in distance is frequent in V-w-5. This is because it gets more difficult to stop, and at that point users pass back and forth across the target if the speed increases.

From these observations, we can see that if the speed increases moderately it takes less time to reach the goal, while it takes more time and more distance if the speed increases excessively. In subjective evaluation, V-w-1 and V-w-2 received good results for ease of operation, while some subjects answered that they become irritated at the slow speed in V-w-1. Therefore, we decided that a mapping between V-w-2 and V-w-3 would be appropriate to shorten the migration time and reduce useless movement.

4 Conclusion

This paper described VisTA and VisTA-walk, which were developed as parts of the Meta-Museum project. With the goal of assisting experts in their studies, VisTA allows them to easily set up and test hypotheses on the space-time transition of ancient villages by visualizing simulated transition processes of the villages with 3D CG. On the other hand, VisTA-walk is a tool for experts to convey their knowledge acquired with VisTA to the general public; furthermore, it has a gesture-based user interface, different from VisTA, that is suitable for use in museum exhibits. We evaluated the effectiveness of the user interface by subjective experi-

[2]The condition to clear each target is to stop within a three-meter radius from the center of the house.

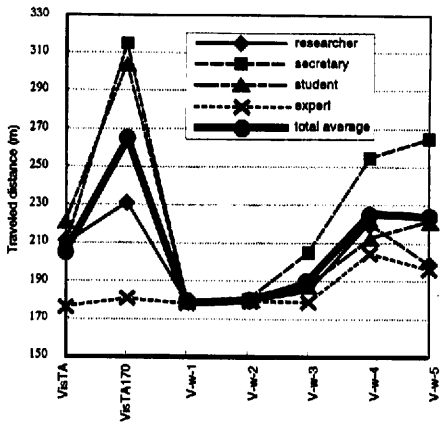

Figure 9. Average traveled distance

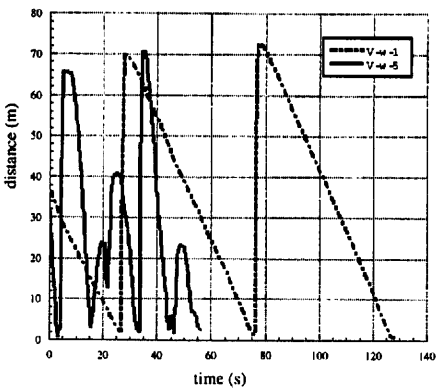

Figure 10. Typical example of movement

ment. The results show that the gesture-based user interface is suitable for museum exhibits because it is easy to use and can give a very realistic sensation without the need for burdensome external devices.

From next spring, VisTA-walk will be displayed in a state-of-the-art technology gallery, which we regard as a predecessor to the museum of the future. We hope to observe how visitors interact with VisTA-walk and then propose better user interfaces based on the results.

Acknowledgments The authors would like to thank Mr. Yasuyoshi Sakai, Dr. Ryohei Nakatsu and the member of the ATR Media Integration & Communications Research Laboratories for the research opportunity and helpful advices. The authors would like to thank Mr. Eduardo Neeter and Mr. Tadashi Takumi for their cooperation in developing VisTA and VisTA-walk, and the Perceptual Computing Section of the MIT Media Lab. for providing the pfinder program.

References

[1] S. Fels, D. Reiners, and K. Mase, "Iamascope: An Interactive Kaleidoscope," *HCI International'97*, San Francisco, 1997.

[2] S. Flavia, "Choreographing media for interactive virtual environments," *Master's thesis, Media Arts and Sciences*, M.I.T., 1996.

[3] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer vision for computer games," *2nd International Conference on Automatic Face and Gesture Recognition*, Killington, VT, USA, pp. 100–105, 1996.

[4] M. Fukumoto, K. Mase, and Y. Suenaga, "Finger-pointer: Pointing interface by image processing," *Comput. & Graphics.*, vol. 18, no. 5, pp. 633–642, 1994.

[5] R. Kadobayashi and K. Mase, "MetaMuseum as A New Communication Environment," *Proc. of Multimedia Communication and Distributed Processing System Workshop*, pp. 71–78, 1995 (in Japanese).

[6] W. A. Kellogg, J. M. Carroll and J. T. Richards, "Making reality a cyberspace," in Cyberspace, M. Benedikt, ed., MIT Press, 1991.

[7] S. Kobayashi, " Optical Gesture Recognition System," *SIGGRAPH97 Visual Proceedings*, Electric Garden, pp.117, 1997.

[8] P. Maes, T. Darrell, B. Blumberg, and A. Pentland, "The ALIVE system: Full-body interaction with autonomous agents," *Proc. of the Computer Animation '95 Conference*, Geneva, Switzerland, 1995.

[9] K. Mase, R. Kadobayashi, and R. Nakatsu, "Meta-museum: A supportive augmented reality environment for knowledge sharing," *Int'l Conf on Virtual Systems and Multimedia '96*, pp. 107–110, Gifu, Japan, 1996.

[10] K. Mase and R. Kadobayashi, "Gesture interface for a virtual walk-through," *Workshop on Perceptual User Interfaces*, pp. 20–21, Banff, 1997.

[11] C. S. Pinhanez, K. Mase, and A. Bobick, "Interval Scripts: a Design Paradigm for Story-Based Interactive Systems," in *CHI97 Conference Proc.*, pp. 287–294, Atlanta, GA, 1997.

[12] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[13] C. R. Wren, F. Sparacino and A. Azarbayejani, T. J. Darrell, T. E. Starner, A. Kotani, C. M. Chao, M. Hlavac, K. B. Russell, and A. Pentland "Perceptive Spaces for Performance and Entertainment: Untethered Interaction using Computer Vision and Audition," *Applied Artificial Intelligence*, vol. 11, no. 4, pp. 267–284, 1997.

[14] Yokohama City Treasure Trove Research Center, "Otsuka Iseki: Kouhoku New Town Excavation Report XII", 1991.

[15] S. Zhai, P. Milgram and D. Drascic, "An evaluation of four 6-degree-of-freedom input techniques," *Adjunct Proc. of INTERCHI'93: ACM Conference on Human Factors in Computing Systems*, 1993.

[16] http://vismod.www.media.mit.edu/vismod/demos/kidsroom/